frontiers in Robotics and AI

PERSPECTIVE
published: 14 February 2019
doi: 10.3389/frobt.2019.00006

Artificial Development by Reinforcement Learning Can Benefit From Multiple Motivations

Günther Palm and Friedhelm Schwenker*

Institute of Neural Information Processing, Ulm University, Ulm, Germany

Research on artificial development, reinforcement learning, and intrinsic motivations like curiosity could profit from the recently developed framework of multi-objective reinforcement learning. The combination of these ideas may lead to more realistic artificial models for life-long learning and goal directed behavior in animals and humans.

Keywords: reinforcement learning, multi-objective, actor-critic design, artificial curiosity, artificial cognition, intrinsic motivation

OPEN ACCESS

Edited by:
Gianluca Baldassarre, Italian National Research Council (CNR), Italy

Reviewed by:
Eiji Uchibe, Advanced Telecommunications Research Institute International (ATR), Japan
Jaan Aru, University of Tartu, Estonia

****Correspondence:***
Günther Palm
guenther.palm@uni-ulm.de

Specialty section:
This article was submitted to Computational Intelligence, a section of the journal Frontiers in Robotics and AI

Received: *30 May 2018*
Accepted: *16 January 2019*
Published: *14 February 2019*

Citation:
Palm G and Schwenker F (2019) Artificial Development by Reinforcement Learning Can Benefit From Multiple Motivations. Front. Robot. AI 6:6. doi: 10.3389/frobt.2019.00006

INTRODUCTION

Reinforcement learning (RL) is a well-established learning paradigm, first consolidated in the book of Sutton and Barto (1998) after the early years of artificial neural networks and machine learning, with strong roots in the mathematics of dynamical programming (Bellman, 1957) and in the early behavioral psychology of Pavlovian conditioning and learning (Rescorla and Wagner, 1972).

In recent years, plausible neural mechanisms for all essential components of RL have been found in the brain, in particular in the basal ganglia, but also in frontal cortical areas, perhaps involved in different versions of RL (Wiering and van Otterlo, 2012), which have been developed not only from a technical, but also from a neuroscientific motivation; overviews are given in Farries and Fairhall (2007), Botvinick et al. (2009), Chater (2009), Maia (2009), Joiner et al. (2017), and Wikenheiser and Schoenbaum (2016).

Also in recent developments of robotics, artificial agents, or artificial life, in particular when the focus is on learning interesting "cognitive" abilities or behaviors or on child-like "artificial development" (Oudeyer et al., 2007), the framework of RL is often used. If it is understood to include its continuous version, actor critic design (Bertsekas and Tsitsiklis, 1996; Prokhorov and Wunsch, 1997) reinforcement learning is a very general approach encompassing applications from Go-playing (Silver et al., 2016) to motor control (Miller et al., 1995; Kretchmara et al., 2001; Todorov, 2004; Schaal and Schweighofer, 2005; Lendaris, 2009; Riedmiller et al., 2009; Wong and Lee, 2010; Little and Sommer, 2011).

Here we are considering RL in the context of robotics or rather of artificial agents that learn to act appropriately in a simulated or real environment. Most often this involves continuous state and action spaces which cannot simply be discretized (Lillicrap et al., 2015). So usually the RL paradigm is combined with a neural network approach to represent the reward predicting function (Sutton and Barto, 1998; Oubbati et al., 2012, 2014; Faußer and Schwenker, 2015).

In this context there are a number of issues that this framework cannot easily accommodate:

1. the learning of several partially incompatible behaviors,
2. the balance between exploration and exploitation,
3. the development and integration of "meta-heurictics" like "curiosity" or "cautiousness,"
4. the problem of finding a "state space" and its partial observability,
5. the simulation of apparently changing strategies in animal behavior.

In reaction to the first issue one might argue that RL is just for one particular behavior, not for the combination of several behaviors; for this one would need to combine several instances of RL. Of course, one could also argue that each animal has just one behavior which maximizes its chance of survival and apparent particular behaviors or motives driving it must be subordinate to this ultimate goal, similarly in economic decision making the ultimate goal is financial utility (money) and it would be irrational to follow other rewards from time to time (as in the fairy tale of *Hans im Gück*). All this has been debated at length (e.g., Simon, 1955, 1991; Tisdell, 1996; Gigerenzer and Selten, 2002; Kahneman, 2003; Dayan and Niv, 2008; Dayan and Seymour, 2009; Glimcher et al., 2009; Chiew and Braver, 2011) leading to considerable doubts in a simple utilitarian view in economy and practically to various approaches extending basic RL, often in a hierarchical fashion (Barto et al., 2004; Botvinick et al., 2009). Even a human or robot Go-player has not only to consider Go strategies, but also (on a lower level) to control his arm movements when taking and placing a piece.

The balance between exploration and exploitation has been widely discussed in classical RL and even before that (e.g., Feldbaum, 1965). It has lead to various, often stochastic, amendments to the original basic method (Wiering and van Otterlo, 2012) without a convincing general solution that works well in most applications. This problem has also inspired more general approaches in more complex scenarios which add special "meta-objectives" like "curiosity" or "cautiousness" to the RL scheme (perhaps first by Schmidhuber, 1991), which again points toward a multi-objective approach. Recently these ideas are discussed in particular in the context of autonomous "cognitive" agents and their "artificial development" (Weng et al., 2001; Lungarella et al., 2003; Barto et al., 2004; Oudeyer et al., 2007).

In biology and human psychology or sociology it is clear that the state space (i.e., the total relevant state of the world) is far from being observable by the senses of the individual animal or human. It might even be doubted whether there is such a state at all. At least it is often asking too much to assume that the individual possesses a representation of the set or space of all possible states. Such scenarios are even outside the usual relatively broad POMDP (partially observable Markov decision process, see Kaelbling et al., 1996) formalism, so biologically motivated realizations of RL often rest on relatively simple versions of RL that don't require knowledge of a "state" in the sense of physics, but just rely on sensory and reward input.

Also the last issue is clearly at variance with the basic model of classical RL. However, when we consider the creation of artificial autonomous agents or artificial animals an obvious potential answer to all of these issues comes to mind: Such an agent or animal usually has several different, sometimes conflicting goals or motivations (e.g., food, drink, and sex) which cannot simply be combined linearly to form one general objective (Liu et al., 2015).

It therefore seems natural to use different instances of RL on different simplified state spaces, which contain incomplete information on different aspects of the physical state of the world, with different objectives or reward functions in different contexts or situations and somehow select the most important ones to determine the agent's behavior in each concrete situation. This means that one has to consider multiple objectives and their interaction in decision making. This problem is studied by a growing research community under the heading of "multiple objective reinforcement learning" (MORL).

The framework of MORL can be used to address and alleviate the 5 problems mentioned above. In fact, it is directly motivated from problems 1 and 5. The dilemma between exploration and exploitation (problem 2) is greatly alleviated by the simple observation that behavior guided by exploitation of one objective usually can be considered as exploration for all other objectives. The development of meta-heuristics or "intrinsic motivations" (issue 3) can be very useful also in technical applications; for the MORL framework advocated here the point is simply to put intrinsic motivations like curiosity or cautiousness side-by-side with the basic "extrinsic" motivation(s). Concerning the state-space (problem 4), in many practical applications a real "state-space" is unknown or at best partially observable. In this case the best one can do is to obtain a sufficiently rich approximate representation for it based on sensory data and reinforcement signals, and more such signals are certainly better than less for this purpose.

REPRESENTING THE STATE SPACE

In order to obtain an approximate state representation by learning from experience, one can use a neural network, typically a multilayer perceptron (MLP) or "deep network" or methods of reservoir computing (Maass et al., 2002; Jaeger and Haas, 2004) for continuous temporal dynamics, or a combination of both. In complex control problems (Koprinkova-Hristova and Palm, 2010) such a representation is often called a "forward model." So the agent (biological or artificial) tries to learn a "state representation network," i.e., a (typically recurrent) network that predicts the next state from a representation of the current state, which integrates sensory input information over time and can be used as input to the evaluation or critic network in the usual situation where the current sensory input is insufficient to determine the "state" of the environment; see for example (Sutton and Barto, 1981; Schmidhuber, 1991; Dayan and Sejnowski, 1996; Herrmann et al., 2000; Gläscher et al., 2010). Such a network can be used as the basis for a second network representing the quality or value function in reinforcement learning or actor-critic design.

The use of neural networks or parameterized approximators as estimators of the state-value or state-action-value function is a way to deal with large or continuous action and state spaces. The approximating function may be a linear or nonlinear function of their parameters, but linear approximators show limitations in their expressive power, while convergence of learning is quaranteed. Nonlinear approximators, typically neural networks, are universal approximators (Cybenko, 1989), but often show instable behavior during learning. During the last years increasingly complex networks are used in RL for large and continuous state spaces; in addition to classical multilayer perceptrons or radial basis function networks, also trainable recurrent neural networks (Hagenbuchner et al., 2017) or echo-state-networks (Scherer et al., 2008; Oubbati et al.,

2012, 2013, 2014; Koprinkova-Hristova et al., 2013) are used, and particular methods have been developed to improve the stability of learning (Hafner and Riedmiller, 2011; Silver et al., 2014; Faußer and Schwenker, 2015; Lillicrap et al., 2015; Parisi et al., 2017). Recently, deep neural networks such as autoencoders and convolutional neural networks have been applied for representation learning and used in combination with RL methods to learn complex decision task from raw data (Lillicrap et al., 2015; Mnih et al., 2015; Mossalam et al., 2016; Srinivasan et al., 2018).

In any case it is practically important for MORL to use one and the same network as a basis to create a sufficiently rich representation in order to train all different objectives (critics and actors) as outputs of the last layer (Mossalam et al., 2016).

Based on the sensory input alone, but also on such an approximate state representation, it often will not be possible to predict the expected reward or the next state with certainty. In a neural network for classification, for example, this uncertainty will be expressed by submaximal activation of several output neurons and these activations may be interpreted as *a posteriori* probabilities of the various outcomes (states or values); the uncertainty in estimating the expected reward is often measured by its variance. Beyond variance, there are various formalisms for calculating measures of certainty or uncertainty from these probabilities, often in terms of information theory (Palm, 2012), and several approaches to incorporate measures of uncertainty, or of "novelty" or "surprise" into the choice of appropriate actions in reinforcement learning (e.g., MacKay, 1992; Sporns and Pegors, 2003; Little and Sommer, 2011; Tishby and Polani, 2011; Sledge and Príncipe, 2017); much of this is reviewed and discussed by Schmidhuber (1997) or Schmidhuber (2003) also in relation to the exploration-exploitation dilemma (Dayan and Sejnowski, 1996; Auer, 2002; Tokic and Palm, 2012; Tokic et al., 2013). Again these practically important considerations point toward MORL, for example in the direction of additional "meta-objectives" like curiosity or cautiousness (Wiering and Schmidhuber, 1998; Uchibe and Doya, 2008; Oubbati et al., 2013). It is often useful to consider at least two versions of the primary objective, namely its expected value and an estimate of the value that can at least be obtained with a reasonably high probability (e.g., the 5-percentile).

The MORL idea transforms the original problem of learning one behavior that is useful in all circumstances into a problem of designing an appropriate architecture for learning and decision making that combines several (probably hierarchically organized) instances or stages of classical RL and possibly other methods of learning or decision making (Oubbati and Palm, 2010).

MULTI-OBJECTIVE REINFORCEMENT LEARNING

A framework for studying these problems in the restricted realm of reinforcement learning, which has recently gained increasing popularity, is called MORL (see Roijers et al., 2013; Liu et al., 2015). We would like to propose to use this framework as a starting point to tackle the broader architectural problem in some concrete scenarios, which occur quite naturally in many technical optimization and control problems and have been elaborated in the MORL community, some examples (Deep Sea Treasure, Bonas World, Cart Pole, Water Reservoir, Resource Gathering, Predator Prey) are described in Drugan et al. (2017) and the literature cited therein; see also Vamplew et al. (2011).

The difference of MORL to classical RL is quite simple: If we think in terms of actor-critic design, where essentially an evaluation of the agent's actions is learned in a POMDP and where this evaluation function may be learned by a neural network, now we just have a vector of evaluations instead of a single value (in the output layer of the network). Similarly there is now an actor for each component of the evaluation vector suggesting an appropriate action for that particular value, objective, or motive. This model clearly leads to the problem how to combine the different objectives and suggested actions in order to decide on the next action. This problem has been discussed thoroughly in the MORL community; for an overview see Liu et al. (2015) and Drugan et al. (2017) and we will contribute a few ideas on this issue in terms of the computational architecture. The most common idea is to combine the different reward values into a weighted sum and take the best action for this combination. More complex methods consider the so-called *pareto-front*, well-known from classical multi-objective optimization. In fact, much of the discussion on optimal decision making for multiple objectives and methods for finding the pareto-optimal solutions (Das and Dennis, 1998; Miettinen, 1999; Mueller-Gritschneder et al., 2009; Motta et al., 2012) can be useful for MORL (see Van Moffaert and Nowé, 2014; Pirotta et al., 2015; Vamplew et al., 2017).

Once the most appropriate action has been determined and carried out, each of the actors and critics is able to learn something from its outcome leading to a modification of the corresponding neural networks, usually through backpropagation of the expected reward update or temporal difference.

From introspection, but also from behavioral animal experiments one gets the impression that each of these motives enters the final evaluation and decision with its own weight or "urgency" that may vary with time, depending on the agent's needs, which implies that there is no fixed "trading relation" between the different motives and their corresponding reward values, so they cannot be reduced to just one value. Modeling artificial agents in this wider framework entails some new problems and tasks, which may also lead to new interesting research projects and interactions with behavioral biologists and psychologists.

Here we describe the basic theoretical framework for this approach:

1. Given n motives, n current predicted values $(v_1, \ldots, v_n)$, and n "urgency weights" $(w_1, \ldots, w_n)$ for them, how do we combine them to one value that should be maximized by the next action? There are different more or less obvious ideas for this (see e.g., Boutilier, 2002; Castelletti et al., 2002; Natarajan and Tadepalli, 2005; Wiering and De Jong, 2007)

also motivated by modeling animal behavior, or reflecting the introspective difference between positive and negative rewards, or between goal seeking and pain avoidance, the most obvious and simple being the weighted sum $v = \sum_i w_i v_i$. At the opposite extreme we would follow the one objective that has maximal $w_i v_i$, or we could consider a minimal value for some objectives as a constraint in maximizing the weighted sum of the others. Here the "higher" motives like curiosity are put side-by-side with "lower" ones like "hunger," which may be psychologically somewhat unsettling, but might actually work. We first encountered this idea in the work of Dörner (2001), see also Bach (2009) and Bach (2012).

2. For each of the motives, in addition to defining the corresponding rewards r_i we have to model their "urgency function" $w_i(t)$. This may involve a dynamical system model of the agent's body and as such may be considered as part of the world model. In particular, it will use the corresponding rewards $r_i(t)$ as inputs. In extreme cases w_i may even be constant or it may simply integrate the incoming rewards as

$$\dot{w}_i(t) = a - b r_i(t) \quad \text{or} \quad \tau \dot{w}_i(t) = -w_i(t) - b r_i(t) + a$$

but much more is easily conceivable, for instance involving thresholds at which the urgency changes drastically. The development of such dynamical models of urgency may be an interesting line of research also in modeling animal behavior. Actually, the simple integration model was probably first introduced informally by Lorenz (1978).

3. It is now possible to introduce some more "cognitive" motives like "curiosity" (see also Pisula, 2009), for which we have to define $r_i(t)$ and $w_i(t)$. For example for curiosity it is natural to define surprising events as rewarding, where surprise may be defined as $-\log p$ relative to a probabilistic world model that the agent may have learnt (Palm, 2012). More concretely, if in world state x the agent receives the observation $o(x)$, or the state description $d(x)$ (Palm, 2012), which has the probability $p(x) = p(d(x))$ in his current model, then his surprise is $-\log p(x)$. Then again $w_i(t)$ can be defined for example by an integration model.
4. Finally we have to decide for the optimal action. Given our estimates for the temporal rewards and urgencies of the different motives and also our momentary combined reward, we can use methods of multi-objective or of plain optimization to find the optimal action. As a starting point we can use the actor outputs for the individual motives and perhaps try their combinations. Practical methods for finding a reasonable solution to the optimization problem in short time are also discussed in the literature on RL and MORL (Handa, 2009; Kooijman et al., 2015; Brys et al., 2017; Parisi et al., 2017; Vamplew et al., 2017).

This leads to an extended RL-architecture, which may be biologically more realistic. Such a more complex architecture also offers interesting additional possibilities for improving behaviors by learning: The existence of more objectives compared to just one, generates a richer representation of (the value of) the current situation, which can be used also to improve the sensory-based world model. It also gives a new perspective on the exploration-exploitation dilemma, since following exploitation of one objective may serve as exploration of the others. We have presented a basic layout of such a multi-objective agent architecture and started some preliminary experiments on it (Oubbati et al., 2013, 2014), but we believe that much more can and should be done in this direction.

AUTHOR CONTRIBUTIONS

GP: involved in preparing the concept of the paper, and writing of the paper; FS: writing of paper including literature work and proofreading.

REFERENCES

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* 3, 397–422.

Bach, J. (2009). *Principles of Synthetic Intelligence*. New York, NY: Oxford University Press.

Bach, J. (2012). A framework for emergent emotions, based on motivation and cognitive modulators. *Int. J. Synthet. Emot.* 3, 43–63. doi: 10.4018/jse.2012010104

Barto, A. G., Singh, S., and Chentanez, N. (2004). "Intrinsically motivated learning of hierarchical collections of skills," in *Proceedings of the 3rd International Conference on Development and Learning* (Cambridge, MA), 112–119.

Bellman, R. E. (1957). *Dynamic Programming*. New York, NY: Dover Publications, Incorporated.

Bertsekas, D. P., and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming, 1st Edn*. Belmont, MA: Athena Scientific.

Botvinick, M. M., Niv, Y., and Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition* 113, 262–280. doi: 10.1016/j.cognition.2008.08.011

Boutilier, C. (2002). "A pomdp formulation of preference elicitation problems," in *AAAI/IAAI* (Edmonton, AB), 239–246.

Brys, T., Harutyunyan, A., Vrancx, P., Nowé, A., and Taylor, M. E. (2017). Multi-objectivization and ensembles of shapings in reinforcement learning. *Neurocomputing* 263, 48–59. doi: 10.1016/j.neucom.2017.02.096

Castelletti, A., Corani, G., Rizzoli, A., Soncini Sessa, R., and Weber, E. (2002). "Reinforcement learning in the operational management of a water system," in *IFAC Workshop on Modeling and Control in Environmental Issues* (Yokohama), 303–308.

Chater, N. (2009). Rational and mechanistic perspectives on reinforcement learning. *Cognition* 113, 350–364. doi: 10.1016/j.cognition.2008.06.014

Chiew, K. S., and Braver, T. S. (2011). Positive affect versus reward: emotional and motivational influences on cognitive control. *Front. Psychol.* 2:279. doi: 10.3389/fpsyg.2011.00279

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Math. Cont. Signals Syst.* 2, 303–314. doi: 10.1007/BF02551274

Das, I., and Dennis, J. E. (1998). Normal-boundary intersection: a new method for generating the Pareto surface in nonlinear multicriteria optimization problems. *SIAM J. Optim.* 8, 631–657. doi: 10.1137/S1052623496307510

Dayan, P., and Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Curr. Opin. Neurobiol.* 18, 185–196. doi: 10.1016/j.conb.2008.08.003

Dayan, P., and Sejnowski, T. J. (1996). Exploration bonuses and dual control. *Mach. Learn.* 25, 5–22. doi: 10.1007/BF00115298

Dayan, P., and Seymour, B. (2009). "Values and actions in aversion," in *Neuroeconomics: Decision Making and the Brain*, eds P. W. Glimcher, C. F.

Camerer, E. Fehr, and R. A. Poldrack (San Diego, CA: Elsevier Academic Press), 175–191.

Dörner, D. (2001). *Bauplan für eine Seele*. Reinbek: Rowohlt.

Drugan, M. M., Wiering, M., Vamplew, P., and Chetty, M. (2017). Special issue on multi-objective reinforcement learning. *Neurocomputing* 263, 1–2. doi: 10.1016/j.neucom.2017.06.020

Farries, M. A. and Fairhall, A. L. (2007). Reinforcement learning with modulated spike timing–dependent synaptic plasticity. *J. Neurophysiol.* 98, 3648–3665. doi: 10.1152/jn.00364.2007

Faußer, S., and Schwenker, F. (2015). Neural network ensembles in reinforcement learning. *Neural Process. Lett.* 41, 55–69. doi: 10.1007/s11063-013-9334-5

Feldbaum, A. (1965). *Optimal Control Systems*. New York, NY: Academic Press.

Gigerenzer, G., and Selten, R. (2002). *Bounded Rationality: The Adaptive Toolbox*. Cambridge: MIT Press.

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66, 585–595. doi: 10.1016/j.neuron.2010.04.016

Glimcher, P. W., Camerer, C. F., Fehr, E., and Poldrack, R. A. (2009). *Introduction: A Brief History of Neuroeconomics*. San Diego, CA: Elsevier Academic Press.

Hafner, R., and Riedmiller, M. (2011). Reinforcement learning in feedback control. *Mach. Learn.* 84, 137–169. doi: 10.1007/s10994-011-5235-x

Hagenbuchner, M., Tsoi, A. C., Scarselli, F., and Zhang, S. (2017). "A fully recursive perceptron network architecture," in *2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017* (Honolulu, HI) , 1–8.

Handa, H. (2009). "Solving multi-objective reinforcement learning problems by eda-rl-acquisition of various strategies," in *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on* (Pisa: IEEE), 426–431.

Herrmann, J. M., Pawelzik, K., and Geisel, T. (2000). Learning predictive representations. *Neurocomputing* 32–33, 785–791. doi: 10.1016/S0925-2312(00)00245-9

Jaeger, H., and Haas, H. (2004). Harnessing nonlinearity: predicting chaotic systems and saving energy in wireless communication. *Science* 304, 78–80. doi: 10.1126/science.1091277

Joiner, J., Piva, M., Turrin, C., and Chang, S. W. (2017). Social learning through prediction error in the brain. *npj Sci. Learn.* 2:8. doi: 10.1038/s41539-017-0009-2

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: a survey. *J. Artif. Intell. Res.* 4, 237–285. doi: 10.1613/jair.301

Kahneman, D. (2003). Maps of bounded rationality: psychology for behavioral economics. *Am. Econ. Rev.* 93, 1449–1475. doi: 10.1257/000282803322655392

Kooijman, C., de Waard, M., Inja, M., Roijers, D. M., and Whiteson, S. (2015). Pareto local policy search for momdp planning. *22th ESANN* (Bruges), 53–58. Available online at: http://www.i6doc.com/en/

Koprinkova-Hristova, P., Oubbati, M., and Palm, G. (2013). Heuristic dynamic programming using echo state network as online trainable adaptive critic. *Int. J. Adapt. Control Signal Process.* 27, 902–914. doi: 10.1002/acs.2364

Koprinkova-Hristova, P., and Palm, G. (2010). "Adaptive critic design with esn critic for bioprocess optimization," in *International Conference on Artificial Neural Networks* (Berlin; Heidelberg: Springer), 438–447.

Kretchmara, R. M., Young, P. M., Anderson, C. W., Hittle, D. C., Anderson, M. L., and Delnero, C. (2001). "Robust reinforcement learning control," in *American Control Conference, 2001. Proceedings of the 2001*, Vol. 2 (Arlington, VA: IEEE), 902–907.

Lendaris, G. G. (2009). "A retrospective on adaptive dynamic programming for control," in *Proceedings of the 2009 International Joint Conference on Neural Networks, IJCNN'09* (Piscataway, NJ: IEEE Press), 945–952.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., et al. (2015). Continuous control with deep reinforcement learning. *arXiv[preprint]. arXiv:1509.02971*.

Little, D. Y., and Sommer, F. T. (2011). Learning in embodied action-perception loops through exploration. *arXiv[preprint]. arXiv:1112.1125*.

Liu, C., Xu, X., and Hu, D. (2015). Multiobjective reinforcement learning: a comprehensive overview. *IEEE Trans. Syst. Man Cybern. Syst.* 45, 385–398. doi: 10.1109/TSMC.2014.2358639

Lorenz, K. (1978). *Vergleichende Verhaltensforschung: Grundlagen der Ethologie*. New York, NY: Springer.

Lungarella, M., Metta, G., Pfeifer, R., and Sandini, G. (2003). Developmental robotics: a survey. *Connect. Sci.* 15, 151–190. doi: 10.1080/09540090310001655110

Maass, W., Natschläger, T., and Markram, H. (2002). Real-time computing without stable states: a new framework for neural computation based on perturbations. *Neural Comput.* 14, 2531–2560. doi: 10.1162/089976602760407955

MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Comput.* 4, 590–604. doi: 10.1162/neco.1992.4.4.590

Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: successes and challenges. *Cogn. Affect. Behav. Neurosci.* 9, 343–364. doi: 10.3758/CABN.9.4.343

Miettinen, K. (1999). *Nonlinear Multiobjective Optimization*. New York, NY: Springer.

Miller, W. T., Sutton, R. S., and Werbos, P. J. (eds.). (1995). *Neural Network and Control*. Cambridge MA: MIT Press.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. (2015). Human-level control through deep reinforcement learning. *Nature* 518:529. doi: 10.1038/nature14236

Mossalam, H., Assael, Y. M., Roijers, D. M., and Whiteson, S. (2016). Multi-objective deep reinforcement learning. *arXiv[preprint]. arXiv:1610.02707*.

Motta, R. S., Afonso, S. M. B., and Lyra, P. R. M. (2012). A modified nbi and nc method for the solution of n-multiobjective optimization problems. *Struct. Multidiscip. Optim.* 46, 239–259. doi: 10.1007/s00158-011-0729-5

Mueller-Gritschneder, D., Graeb, H., and Schlichtmann, U. (2009). A successive approach to compute the bounded Pareto front of practical multiobjective optimization problems. *SIAM J. Optim.* 20, 915–934. doi: 10.1137/080729013

Natarajan, S., and Tadepalli, P. (2005). "Dynamic preferences in multi-criteria reinforcement learning," in *Proceedings of the 22nd International Conference on Machine Learning* (Bonn: ACM), 601–608.

Oubbati, M., Kord, B., Koprinkova-Hristova, P., and Palm, G. (2014). Learning of embodied interaction dynamics with recurrent neural networks: some exploratory experiments. *J. Neural Eng.* 11:026019. doi: 10.1088/1741-2560/11/2/026019

Oubbati, M., Oess, T., Fischer, C., and Palm, G. (2013). "Multiobjective reinforcement learning using adaptive dynamic programming and reservoir computing," in *Re- inforcement Learning with Generalized Feedback: Beyond Numeric Rewards (ECML 2013)* (Prague).

Oubbati, M., and Palm, G. (2010). A neural framework for adaptive robot control. *Neural Comput. Appl.* 19, 103–114. doi: 10.1007/s00521-009-0262-2

Oubbati, M., Uhlemann, J., and Palm, G. (2012). "Adaptive learning in continuous environment using actor-critic design and echo-state networks," in *International Conference on Simulation of Adaptive Behavior* (Berlin; Heidelberg: Springer), 320–329.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Trans. Evol. Comput.* 11, 265–286. doi: 10.1109/TEVC.2006.890271

Palm, G. (2012). *Novelty, Information and Surprise*. Heidelberg; New York, NY; Dordrecht; Berlin: Springer Science & Business Media.

Parisi, S., Pirotta, M., and Peters, J. (2017). Manifold-based multi-objective policy search with sample reuse. *Neurocomputing* 263, 3–14. doi: 10.1016/j.neucom.2016.11.094

Pirotta, M., Parisi, S., and Restelli, M. (2015). "Multi-objective reinforcement learning with continuous pareto frontier approximation," in *29th AAAI Conference on Artificial Intelligence, AAAI 2015 and the 27th Innovative Applications of Artificial Intelligence Conference, IAAI 2015* (Austin: AAAI Press), 2928–2934.

Pisula, W. (2009). *Curiosity and Information Seeking in Animal and Human Behavior*. Boca Raton, FL: Brown Walker Press.

Prokhorov, D. V., and Wunsch, D. C. (1997). Adaptive critic designs. *IEEE Trans. Neural Netw.* 8, 997–1007. doi: 10.1109/72.623201

Rescorla, R., and Wagner, A. (1972). "A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement," in *Classical Conditioning II: Current Research and Theory*, eds A. Black and W. Prokasy (New York, NY: Appleton Century Crofts), 64–99.

Riedmiller, M., Gabel, T., Hafner, R., and Lange, S. (2009). Reinforcement learning for robot soccer. *Auton. Robots* 27, 55–73. doi: 10.1007/s10514-009-9120-4

Roijers, D. M., Vamplew, P., Whiteson, S., and Dazeley, R. (2013). A survey of multi-objective sequential decision-making. *J. Artif. Int. Res.* 48, 67–113. doi: 10.1613/jair.3987

Schaal, S., and Schweighofer, N. (2005). Computational motor control in humans and robots. *Curr. Opin. Neurobiol.* 15, 675–682. doi: 10.1016/j.conb.2005.10.009

Scherer, S., Oubbati, M., Schwenker, F., and Palm, G. (2008). "Real-time emotion recognition from speech using echo state networks," in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition* (Heidelberg; Berlin: Springer), 205–216.

Schmidhuber, J. (1991). "A possibility for implementing curiosity and boredom in model-building neural controllers," in *Proceedings of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, eds J. A. Meyer and S. W. Wilson (Cambridge, MA: MIT Press), 222–227.

Schmidhuber, J. (1997). *What's Interesting?* Technical Report 35-97, IDSIA.

Schmidhuber, J. (2003). "Exploring the predictable," in *Advances in Evolutionary Computing*, eds S. Ghosh and S. Tsuitsui (Cham: Springer), 579–612.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature* 529, 484–489. doi: 10.1038/nature16961

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014). "Deterministic policy gradient algorithms," in *Proceedings of the 31 International Conference on Machine Learning* (Beijing).

Simon, H. A. (1955). A behavioral model of rational choice. *Q. J. Econ.* 69, 99–118. doi: 10.2307/1884852

Simon, H. A. (1991). Bounded rationality and organizational learning. *Organ. Sci.* 2, 125–134. doi: 10.1287/orsc.2.1.125

Sledge, I. J., and Príncipe, J. C. (2017). "Balancing exploration and exploitation in reinforcement learning using a value of information criterion," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on* (New Orleans, LA: IEEE), 2816–2820.

Sporns, O., and Pegors, T. K. (2003). "Information-theoretical aspects of embodied artificial intelligence," in *Embodied Artificial Intelligence, Volume 2865 of Lecture Notes in Computer Science* (Berlin; Heidelberg: Springer), 74–85.

Srinivasan, S., Lanctot, M., Zambaldi, V., Pérolat, J., Tuyls, K., Munos, R., et al. (2018). "Actor-critic policy optimization in partially observable multiagent environments," in *Advances in Neural Information Processing Systems* (Montreal, QC), 3426–3439.

Sutton, R. S., and Barto, A. G. (1981). An adaptive network that constructs and uses an internal model of its world. *Cogn. Brain Theory* 4, 217–246.

Sutton, R. S., and Barto, A. G. (1998). *Introduction to Reinforcement Learning, 1st Edn*. Cambridge, MA: MIT Press.

Tisdell, C. (1996). *Bounded Rationality and Economic Evolution*. Michigan: Edward Elgar Publishing.

Tishby, N., and Polani, D. (2011). "Information theory of decisions and actions," in *Perception-action Cycle*, eds V. Cutsuridis and A. Hussain, and J. G. Taylor (New York, NY: Springer), 601–636.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nat. Neurosci.* 7:907. doi: 10.1038/nn1309

Tokic, M., and Palm, G. (2012). "Adaptive exploration using stochastic neurons," in *International Conference on Artificial Neural Networks* (Berlin; Heidelberg: Springer), 42–49.

Tokic, M., Schwenker, F., and Palm, G. (2013). "Meta-learning of exploration and exploitation parameters with replacing eligibility traces," in *IAPR International Workshop on Partially Supervised Learning* (Berlin; Heidelberg: Springer), 68–79.

Uchibe, E., and Doya, K. (2008). Finding intrinsic rewards by embodied evolution and constrained reinforcement learning. *Neural Netw.* 21, 1447–1455. doi: 10.1016/j.neunet.2008.09.013

Vamplew, P., Dazeley, R., Berry, A., Issabekov, R., and Dekker, E. (2011). Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Mach. Learn.* 84, 51–80. doi: 10.1007/s10994-010-5232-5

Vamplew, P., Issabekov, R., Dazeley, R., Foale, C., Berry, A., Moore, T., et al. (2017). Steering approaches to Pareto-optimal multiobjective reinforcement learning. *Neurocomputing* 263, 26–38. doi: 10.1016/j.neucom.2016.08.152

Van Moffaert, K., and Nowé, A. (2014). Multi-objective reinforcement learning using sets of Pareto dominating policies. *J. Mach. Learn. Res.* 15, 3483–3512.

Weng, J., McClelland, J., Pentland, A., Sporns, O., Stockman, I., Sur, M., et al. (2001). Autonomous mental development by robots and animals. *Science* 291, 599–600. doi: 10.1126/science.291.5504.599

Wiering, M., and Schmidhuber, J. (1998). "Efficient model-based exploration," in *Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, Vol. 6, (Cambridge, MA: MIT Press), 223–228.

Wiering, M., and van Otterlo, M. (2012). *Reinforcement Learning: State of the Art*. Heidelberg; New York, NY; Dordrecht: Springer. doi: 10.1007/978-3-642-27645-3

Wiering, M. A., and De Jong, E. D. (2007). "Computing optimal stationary policies for multi-objective markov decision processes," in *Approximate Dynamic Programming and Reinforcement Learning, 2007. ADPRL 2007. IEEE International Symposium on* (Honolulu, HI: IEEE), 158–165. doi: 10.1109/ADPRL.2007.368183

Wikenheiser, A. M., and Schoenbaum, G. (2016). Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. *Nat. Rev. Neurosci.* 17, 513–523. doi: 10.1038/nrn.2016.56

Wong, W. C., and Lee, J. H. (2010). A reinforcement learning-based scheme for direct adaptive optimal control of linear stochastic systems. *Opt. Cont. Appl. Methods* 31, 365–374. doi: 10.1002/oca.915

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.